BSN: Boundary Sensitive Network for Temporal Action Proposal Generation

BSN：用于时间行动提案生成的边界敏感网络

日期：2018-09-26

作者：Tianwei Lin (/search?search_txt=Tianwei Lin) 、Xu Zhao (/search?search_txt=Xu Zhao) 、Haisheng Su (/search?search_txt=Haisheng Su) 、Chongjing Wang (/search?search_txt=Chongjing Wang) 、Ming Yang (/search?search_txt=Ming Yang)

论文：http://arxiv.org/pdf/1806.02964v3.pdf (http://arxiv.org/pdf/1806.02964v3.pdf)

Abstract
摘要

Temporal action proposal generation is an important yet challenging problem, since temporal proposals with rich action content are indispensable for analysing real-world videos with long duration and high proportion irrelevant content. This problem requires methods not only generating proposals with precise temporal boundaries, but also retrieving proposals to cover truth action instances with high recall and high overlap using relatively fewer proposals. To address these difficulties, we introduce an effective proposal generation method, named Boundary-Sensitive Network (BSN), which adopts “local to global” fashion. Locally, BSN first locates temporal boundaries with high probabilities, then directly combines these boundaries as proposals. Globally, with Boundary-Sensitive Proposal feature, BSN retrieves proposals by evaluating the confidence of whether a proposal contains an action within its region. We conduct experiments on two challenging datasets: ActivityNet-1.3 and THUMOS14, where BSN outperforms other state-of-the-art temporal action proposal generation methods with high recall and high temporal precision. Finally, further experiments demonstrate that by combining existing action classifiers, our method significantly improves the state-of-the-art temporal action detection performance.

由于具有丰富动作内容的时间提议对于分析具有长持续时间和高比例无关内容的真实世界视频是必不可少的，因此时间动作提议生成是一个重要且具有挑战性的问题。这个问题要求方法不仅要生成具有精确时间边界的提议，而且要求使用相对较少的提案检索提议以涵盖具有高召回率和高重叠度的真实行为实例。为了解决这些困难，我们引入了一种有效的提议生成方法，称为边界敏感网络（BSN），采用“本地到全球”的方式。在本地，BSN首先以高概率定位时间边界，然后直接将这些边界组合为提议。在全球范围内，通过边界敏感提案功能，BSN通过评估提案是否在其区域内包含某个操作的能力来检索提案。我们对两个具有挑战性的数据集进行了实验：ActivityNet-1.3和THUMOS14，其中BSN优于其他具有高召回率和高时间精度的最先进的时间动作建议生成方法。最后，进一步的实验表明，通过结合现有的动作分类器，我们的方法显着改善了最先进的时间动作检测性能。

Keywords: Temporal action proposal generation · Temporal action detection · Temporal convolution · Untrimmed video

关键词：时间行动建议生成·时间动作检测·时间卷积·未修剪的视频

1 Introduction
1简介

Nowadays, with fast development of digital cameras and Internet, the number of videos is continuously booming, making automatic video content analysis methods widely required. One major branch of video analysis is action recognition, which aims to classify manually trimmed video clips containing only one action instance. However, videos in real scenarios are usually long, untrimmed and contain multiple action instances along with irrelevant contents. This problem requires algorithms for another challenging task: temporal action detection, which aims to detect action instances in untrimmed video including both temporal boundaries and action classes. It can be applied in many areas such as video recommendation and smart surveillance.

如今，随着数码相机和互联网的快速发展，视频的数量不断增加，因此广泛需要自动视频内容分析方法。视频分析的一个主要分支是动作识别，其目的是对仅包含一个动作实例的手动修剪视频剪辑进行分类。但是，真实场景中的视频通常很长，未经修剪，并且包含多个动作实例以及不相关的内容。该问题需要用于另一个具有挑战性的任务的算法：时间动作检测，其旨在检测未修剪视频中的动作实例，包括时间边界和动作类。它可以应用于许多领域，如视频推荐和智能监控。

Similar with object detection in spatial domain, temporal action detection task can be divided into two stages: proposal and classification. Proposal generation stage aims ⋆ Corresponding author.

与空间域中的对象检测类似，时间动作检测任务可以分为两个阶段：提议和分类。提案生成阶段的目标⋆通讯作者。

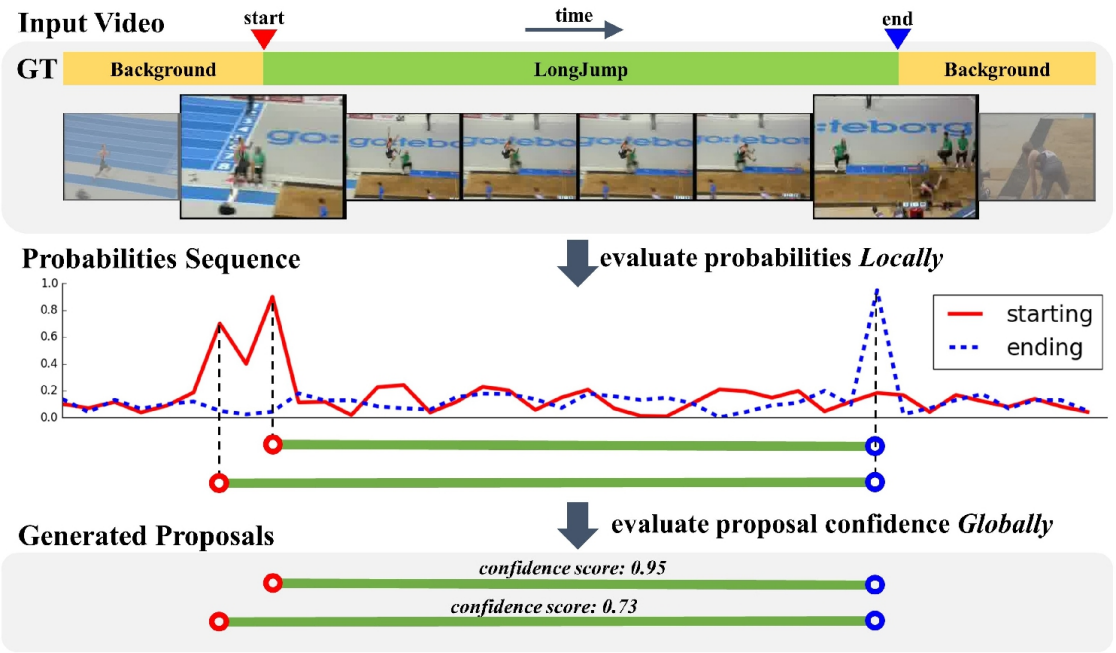


Fig. 1: Overview of our approach. Given an untrimmed video, (1) we evaluate boundaries and actionness probabilities of each temporal location and generate proposals based on boundary probabilities, and (2) we evaluate the confidence scores of proposals with proposal-level feature to get retrieved proposals.

图1：我们的方法概述。给定未修剪的视频，（1）我们评估每个时间位置的边界和行动概率并基于边界概率生成

提议，以及（2）我们评估具有提议级特征的提案的置信分数以获得检索的提议。

to generate temporal video regions which may contain action instances, and classification stage aims to classify classes of candidate proposals. Although classification methods have reached convincing performance, the detection precision is still low in many benchmarks [1,2]. Thus recently temporal action proposal generation has received much attention [3,4,5,6], aiming to improve the detection performance by improving the quality of proposals. High quality proposals should come up with two key properties: (1) proposals can cover truth action regions with both high recall and high temporal overlap, (2) proposals are retrieved so that high recall and high overlap can be achieved using fewer proposals to reduce the computation cost of succeeding steps.
生成可能包含动作实例的时间视频区域，并且分类阶段旨在对候选提议的类别进行分类。虽然分类方法已达到令人信服的性能，但在许多基准测试中检测精度仍然很低[1,2]。因此，最近时间行动提案的产生受到了很多关注[3,4,5,6]，旨在通过提高提案的质量来提高检测性能。高质量的提案应该提出两个关键属性：（1）提案可以涵盖具有高召回率和高时间重叠的真实行动区域。（2）检索提案，以便使用较少的减少提案来实现高召回和高重叠后续步骤的计算成本。

To achieve high proposal quality, a proposal generation method should generate proposals with flexible temporal durations and precise temporal boundaries, then retrieve proposals with reliable confidence scores, which indicate the probability of a proposal containing an action instance. Most recently proposal generation methods [3,4,5,7] generate proposals via sliding temporal windows of multiple durations in video with regular interval, then train a model to evaluate the confidence scores of generated proposals for proposals retrieving, while there is also method [6] making external boundaries regression. However, proposals generated with pre-defined durations and intervals may have some major drawbacks: (1) usually not temporally precise; (2) not flexible enough to cover variable temporal durations of ground truth action instances, especially when the range of temporal durations is large.
为了实现较高的提议质量，提案生成方法应生成具有灵活的时间持续时间和精确的时间边界的提议，然后检索具有可靠的信任分数的提议，其指示包含动作实例的提议的概率。最近的提议生成方法[3,4,5,7]通过定期间隔的视频中多个持续时间的滑动时间窗生成提议，然后训练模型来评估生成的提案检索提案的信度分数，同时还有方法[6]使外部边界回归。但是，使用预先定义的持续时间和间隔生成的建议可能存在一些主要缺点：（1）通常不是在时间上精确; （2）不足以覆盖地面实况动作实例的可变时间持续时间，特别是当时间持续时间范围很大时。

To address these issues and generate high quality proposals, we propose the BoundarySensitive Network (BSN), which adopts “local to global” fashion to locally combine high probability boundaries as proposals and globally retrieve candidate proposals using proposal-level feature as shown in Fig 1. In detail, BSN generates proposals in three steps. First, BSN evaluates the probabilities of each temporal location in video whether it is inside or outside, at or not at the boundaries of ground truth action instances, to generate starting, ending and actionness probabilities sequences as local information.
为了解决这些问题并生成高质量的提案，我们提出了边界敏感网络（BSN），它采用“本地到全球”的方式，在本地将高概率边界组合为提案，并使用提案级功能全局检索候选提案，如图1所示。具体而言，BSN分三步产生提案。首先，BSN评估视频中的每个时间位置的概率，无论其是在地面实况动作实例的边界内部还是外部，在或不在地面实况动作实例的边界处，以生成起始，结束和动作概率序列作为本地信息。

Second, BSN generates proposals via directly combining temporal locations with high starting and ending probabilities separately. Using this bottom-up fashion, BSN can generate proposals with flexible durations and precise boundaries. Finally, using features composed by actionness scores within and around proposal, BSN retrieves proposals by evaluating

the confidence of whether a proposal contains an action. These proposal-level features offer global information for better evaluation. In summary, the main contributions of our work are three-folds: (1) We introduce a new architecture (BSN) based on “local to global” fashion to generate high quality temporal action proposals, which locally locates high boundary probability locations to achieve precise proposal boundaries and globally evaluates proposal-level feature to achieve reliable proposal confidence scores for retrieving.
其次，BSN通过直接组合具有高起始概率和结束概率的时间位置来生成提议。使用这种自下而上的方式，BSN可以生成具有灵活持续时间和精确边界的提议。最后，使用由提案内部和周围的操作性分数组成的功能，BSN通过评估提案是否包含操作的信心来检索提案。这些提案级功能提供全球信息以便更好地进行评估。总之，我们工作的主要贡献有三方面：（1）我们引入了一种基于“局部到全球”方式的新架构（BSN）来生成高质量的时间动作建议，这些建议在本地定位高边界概率位置以实现精确的提案边界并全局评估提案级别功能，以获得可靠的提案保证分数以进行检索。

(2) Extensive experiments demonstrate that our method achieves significantly better proposal quality than other state-of-the-art proposal generation methods, and can generate proposals in unseen action classes with comparative quality.
（2）广泛的实验表明，我们的方法比其他最先进的提案生成方法实现了更好的提案质量，并且可以在具有相对质量的看不见的行动类中产生提议。

(3) Integrating our method with existing action classifier into detection framework leads to significantly improved performance on temporal action detection task.
（3）将我们的方法与现有的动作分类器集成到检测框架中，可以显着提高时间动作检测任务的性能。

2 Related work
2相关工作

Action recognition. Action recognition is an important branch of video related research areas and has been extensively studied. Earlier methods such as improved Dense Trajectory (iDT) [8,9] mainly adopt hand-crafted features such as HOF, HOG and MBH. In recent years, convolutional networks are widely adopted in many works [10,11,12,13] and have achieved great performance. Typically, two-stream network [10,11,13] learns appearance and motion features based on RGB frame and optical flow field separately. C3D network [12] adopts 3D convolutional layers to directly capture both appearance and motion features from raw frames volume. Action recognition models can be used for extracting frame or snippet level visual features in long and untrimmed videos.
行动认可。行动识别是视频相关研究领域的重要分支，并得到了广泛的研究。早期的方法，如改进的密集轨迹（iDT）[8,9]主要采用手工制作的功能，如HOF，HOG和MBH。近年来，卷积网络在许多作品中被广泛采用[10,11,12,13]并取得了很好的表现。通常，双流网络[10,11,13]分别基于RGB帧和光流场学习外观和运动特征。C3D网络[12]采用3D卷积层直接捕获原始帧体积的外观和运动特征。动作识别模型可用于提取长和未修剪视频中的帧或片段级视觉特征。

Object detection and proposals. Recent years, the performance of object detection has been significantly improved with deep learning methods. R-CNN [14] and its variations [15,16] construct an important branch of object detection methods, which adopt “detection by classifying proposals” framework. For proposal generation stage, besides sliding windows [17], earlier works also attempt to generate proposals by exploiting low-level cues such as HOG and Canny edge [18,19]. Recently some methods [16,20,21] adopt deep learning model to generate proposals with faster speed and

stronger modelling capacity. In this work, we combine the properties of these methods via evaluating boundaries and actionness probabilities of each location using neural network and adopting “local to global” fashion to generate proposals with high recall and accuracy.
对象检测和提议。近年来，通过深度学习方法，物体检测的性能得到了显着提高。R-CNN [14]及其变体[15,16]构建了一个重要的目标检测方法分支，采用“分类提案检测”框架。对于提案生成阶段，除了滑动窗口[17]之外，早期的工作还试图通过利用HOG和Canny边缘等低级线索来生成提议[18,19]。最近，一些方法[16,20,21]采用深度学习模型来生成速度更快，建模能力更强的建议。在这项工作中，我们通过使用神经网络评估每个位置的边界和行动概率并采用“局部到全局”方式来生成具有高召回率和准确性的建议，从而结合这些方法的属性。

Boundary probabilities are also adopted in LocNet [22] for revising the horizontal and vertical boundaries of existing proposals. Our method differs in (1) BSN aims to generate while LocNet aims to revise proposals and (2) boundary probabilities are calculated repeatedly for all boxes in LocNet but only once for a video in BSN.
LocNet [22]也采用边界概率来修订现有提案的横向和纵向边界。我们的方法的不同之处在于：（1）BSN的目的是在LocNet旨在修改提案的同时生成;（2）对LocNet中的所有框重复计算边界概率，但对于BSN中的视频仅重复计算一次。

Temporal action detection and proposals. Temporal action detection task aims to detect action instances in untrimmed videos including temporal boundaries and action classes, and can be divided into proposal and classification stages. Most detection methods [7,23,24] take these two stages separately, while there is also method [25,26] taking these two stages jointly. For proposal generation, earlier works [27,28,29] directly use sliding windows as proposals. Recently some methods [3,4,5,6,7] generate proposals with pre-defined temporal durations and intervals, and use multiple methods to evaluate the confidence score of proposals, such as dictionary learning [4] and recurrent neural network [5]. TAG method [24] adopts watershed algorithm to generate proposals with flexible boundaries and durations in local fashion, but without global proposal-level confidence evaluation for retrieving. In our work, BSN can generate proposals with flexible boundaries meanwhile reliable confidence scores for retrieving.
时间行动检测和建议。时间动作检测任务旨在检测未修剪视频中的动作实例，包括时间边界和动作类，并且可以分为提议阶段和分类阶段。大多数检测方法[7,23,24]分别采用这两个阶段，同时还有方法[25,26]联合采用这两个阶段。对于提案生成，早期的工作[27,28,29]直接使用滑动窗口作为提议。最近，一些方法[3,4,5,6,7]生成了具有预定义时间和间隔的提议，并使用多种方法来评估提案的信度得分，如字典学习[4]和递归神经网络[5]]。TAG方法[24]采用分水岭算法以局部方式生成具有灵活边界和持续时间的提议，但没有用于检索的全局提案级别信任评估。在我们的工作中，BSN可以生成具有灵活边界的建议，同时可以获得可靠的检索分数。

Recently temporal action detection method [30] detects action instances based on class-wise start, middle and end probabilities of each location. Our method is superior than [30] in two aspects: (1) BSN evaluates probabilities score using temporal convolution to better capture temporal information and (2) “local to global” fashion adopted in BSN brings more precise boundaries and better retrieving quality.
最近，时间动作检测方法[30]基于每个位置的分类开始，中间和结束概率来检测动作实例。我们的方法在两个方面优于[30]：（1）BSN使用时间卷积评估概率得分以更好地捕获时间信息;（2）BSN中采用的“局部到全局”方式带来更精确的边界和更好的检索质量。

3 Our Approach

3我们的方法

3.1 Problem Definition

3.1问题定义

An untrimmed video sequence can be denoted as $X = \{x_n\}_{n=1}^{l_v}$ with l_v frames, where x_n is the n-th frame in X. Annotation of video X is composed by a set of action instances $\Psi_g = \{\varphi_n = (t_{s,n}, t_{e,n})\}_{n=1}^{N_g}$, where N_g is the number of truth action instances in video X, and $t_{s,n}$, $t_{e,n}$ are starting and ending time of action instance φ_n separately. Unlike detection task, classes of action instances are not considered in temporal action proposal generation. Annotation set Ψ_g is used during training. During prediction, generated proposals set Ψ_p should cover Ψ_g with high recall and high temporal overlap.

未修剪的视频序列可以用l_v帧表示为$X = \{x_n\}_{n=1}^{l_v}$，其中x_n是X中的第n帧。视频X的注释由一组动作实例$\Psi_g = \{\varphi_n = (t_{s,n}, t_{e,n})\}_{n=1}^{N_g}$组成，其中$N_g$是视频X中的真实动作实例的数量，$t_{s,n}$，$t_{e,n}$分别是动作实例$\varphi_n$的开始和结束时间。与检测任务不同，在时间操作提议生成中不考虑动作实例类。训练期间使用注释集Ψ_g。在预测期间，生成的提议集Ψ_p应涵盖具有高召回率和高时间重叠的Ψ_g。

3.2 Video Features Encoding

3.2视频功能编码

To generate proposals of input video, first we need to extract feature to encode visual content of video. In our framework, we adopt two-stream network [11] as visual encoder, since this architecture has shown great performance in action recognition task

为了生成输入视频的提议，首先我们需要提取特征来编码视频的可视内容。在我们的框架中，我们采用双流网络[11]作为可视化编码器，因为这种架构在动作识别任务中表现出了很好的性能

[31] and has been widely adopted in temporal action detection and proposal generation tasks [24,25,32]. Two-stream network contains two branches: spatial network operates on single RGB frame to capture appearance feature, and temporal network operates on stacked optical flow field to capture motion information.

[31]并已在时间动作检测和提议生成任务中被广泛采用[24,25,32]。双流网络包含两个分支：空间网络在单个RGB帧上操作以捕获外观特征，并且时间网络在堆叠光流场上操作以捕获运动信息。

To extract two-stream features, as shown in Fig 2(a), first we compose a snippets sequence $S = \{s_n\}_{n=1}^{l_s}$ from video X, where l_s is the length of snippets sequence. A snippet $s_n = (x_{t_n}, o_{t_n})$ includes two parts: x_{t_n} is the t_n -th RGB frame in X and o_{t_n} is stacked optical flow field derived around center frame x_{t_n}. To reduce the computation cost, we extract snippets with a regular frame interval σ, therefore $l_s = l_v/\sigma$. Given a snippet s_n, we concatenate output scores in top layer of both spatial and temporal networks to form the encoded feature vector $f_{t_n} = (f_{S,t_n}, f_{T,t_n})$, where f_{S,t_n}, f_{T,t_n} are output scores from spatial and temporal networks separately. Thus given a snippets sequence S with length l_s, we can extract a feature sequence $F = \{f_{t_n}\}_{n=1}^{l_s}$. These two-stream feature sequences are used as the input of BSN.

为了提取双流特征，如图2（a）所示，首先我们从视频X组成片段序列$S = \{s_n\}_{n=1}^{l_s}$，其中l_s是片段序列的长度。片段$s_n = (x_{t_n}, o_{t_n})$包括两部分：x_{t_n}是X中的t_n-RGB RGB帧，o_{t_n}是围绕中心帧x_{t_n}导出的堆叠光流场。为了降低计算成本，我们提取具有规则帧间隔σ的片段，因此$l_s = l_v/\sigma$。给定一个片段s_n，我们在空间和时间网络的顶层连接输出分数以形成编码特征向量$f_{t_n} = (f_{S,t_n}, f_{T,t_n})$，其中$f_{S,t_n}$，$f_{T,t_n}$分别从空间和时间网络输

出分数。因此，给定长度为l_s的片段序列S，我们可以提取特征序列$F = \{f_{t_n}\}_{n=1}^{l_s}$。这些双流特征序列用作BSN的输入。

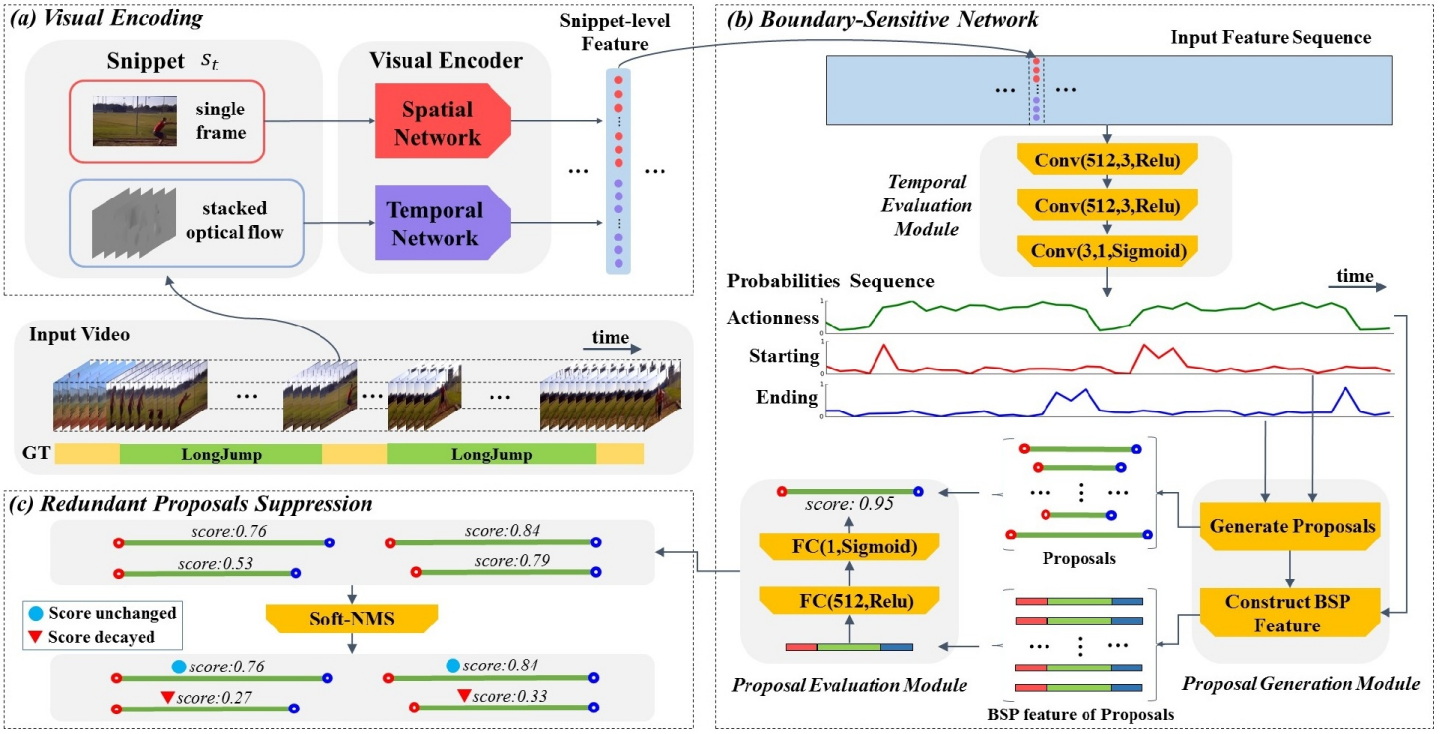


Fig. 2: The framework of our approach. (a) Two-stream network is used for encoding visual features in snippet-level. (b) The architecture of Boundary-Sensitive Network: temporal evaluation module handles the input feature sequence, and evaluates starting, ending and actionness probabilities of each temporal location; proposal generation module generates proposals with high starting and ending probabilities, and construct Boundary-Sensitive Proposal (BSP) feature for each proposal; proposal evaluation module evaluates confidence score of each proposal using BSP feature. (c) Finally, we use Soft-NMS algorithm to suppress redundant proposals by decaying their scores.

图2：我们方法的框架。 （a）双流网络用于编码片段级的视觉特征。 （b）边界敏感网络的体系结构：时间评估模块处理输入特征序列，并评估每个时间位置的起始，结束和动作概率;提议生成模块生成具有高起始和结束概率的提议，并为每个提议构建边界敏感提议（BSP）特征;提案评估模块使用BSP功能评估每个提案的信度得分。 （c）最后，我们使用Soft-NMS算法通过衰减其分数来抑制冗余提议。

3.3 Boundary-Sensitive Network

3.3边界敏感网络

To achieve high proposal quality with both precise temporal boundaries and reliable confidence scores, we adopt “local to global” fashion to generate proposals. In BSN, we first generate candidate boundary locations, then combine these locations as proposals and evaluate confidence score of each proposal with proposal-level feature.

为了通过精确的时间边界和可靠的信心分数实现高提案质量，我们采用“本地到全球”的方式来生成提案。在BSN中，我们首先生成候选边界位置，然后将这些位置组合为提案，并使用提案级别功能评估每个提案的置信度得分。

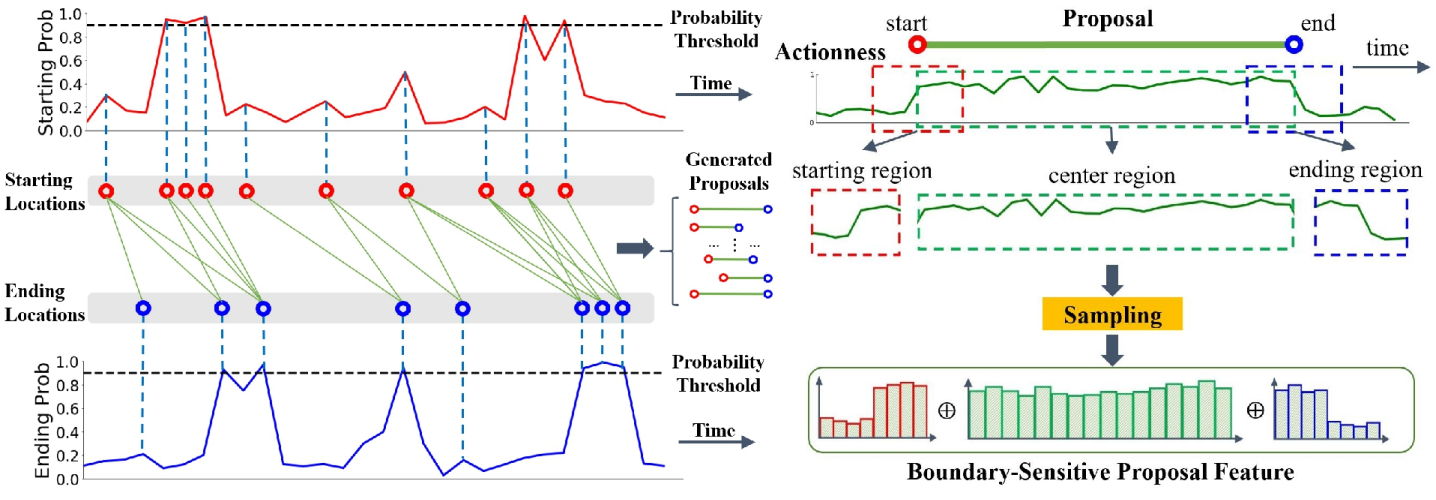
Network architecture. The architecture of BSN is presented in Fig 2(b), which contains three modules: temporal evaluation, proposal generation and proposal evaluation. Temporal evaluation module is a three layers temporal

convolutional neural network, which takes the two-stream feature sequences as input, and evaluates probabilities of each temporal location in video whether it is inside or outside, at or not at boundaries of ground truth action instances, to generate sequences of starting, ending and actionness probabilities respectively. Proposal generation module first combines the temporal locations with separately high starting and ending probabilities as candidate proposals, then constructs Boundary-Sensitive Proposal (BSP) feature for each candidate proposal based on actionness probabilities sequence. Finally, proposal evaluation module, a multilayer perceptron model with one hidden layer, evaluates the confidence score of each candidate proposal based on BSP feature. Confidence score and boundary probabilities of each proposal are fused as the final confidence score for retrieving.

网络架构。 BSN的体系结构如图2（b）所示，其中包含三个模块：时间评估，提议生成和提议评估。时间评估模块是三层时间卷积神经网络，它将两个流特征序列作为输入，并评估视频中每个时间位置的概率，无论它是在地面实况动作实例的边界内部还是外部，分别生成起始，结束和动作概率的序列。提议生成模块首先将时间位置与单独的高起始概率和结束概率组合作为候选提议，然后基于动作概率序列为每个候选提议构建边界敏感提议（BSP）特征。最后，提议评估模块，具有一个隐藏层的多层感知器模型，基于BSP特征评估每个候选提议的信度得分。每个提案的置信分数和边界概率被融合为检索的最终置信分数。

Temporal evaluation module. The goal of temporal evaluation module is to evaluate starting, ending and actionness probabilities of each temporal location, where three binary classifiers are needed. In this module, we adopt temporal convolutional layers upon Fig. 3: Details of proposal generation module. (a) Generate proposals. First, to generate candidate boundary locations, we choose temporal locations with high boundary probability or being a probability peak. Then, we combine candidate starting and ending locations as proposals when their duration satisfying condition. (b) Construct BSP feature. Given a proposal and actionness probabilities sequence, we can sample actionness sequence in starting, center and ending regions of proposal to construct BSP feature.

时间评估模块。时间评估模块的目标是评估每个时间位置的起始，结束和动作概率，其中需要三个二进制分类器。在这个模块中，我们采用时间卷积层，如图3所示：提案生成模块的细节。 （a）提出提案。首先，为了生成候选边界位置，我们选择具有高边界概率或是概率峰值的时间位置。然后，我们将候选起始位置和结束位置组合为其持续时间满足条件的建议。 （b）构建BSP功能。给定提议和动作概率序列，我们可以在提议的起始区域，中心区域和结束区域中采样动作序列以构建BSP特征。



(a) Generate proposals (b) Construct BSP feature

feature sequence, with good modelling capacity to capture local semantic information such as boundaries and actionness

probabilities.
特征序列，具有良好的建模能力，可捕获局部语义信息，如边界和动作概率。

A temporal convolutional layer can be simply denoted as $Conv(c_f, c_k, Act)$, where c_f, c_k and Act are filter numbers, kernel size and activation function of temporal convolutional layer separately. As shown in Fig 2(b), the temporal evaluation module can be defined as $Conv(512, 3, Relu) \rightarrow Conv(512, 3, Relu) \rightarrow Conv(3, 1, Sigmoid)$, where the three layers have same stride size 1. Three filters with sigmoid activation in the last layer are used as classifiers to generate starting, ending and actionness probabilities separately. For convenience of computation, we divide feature sequence into non-overlapped windows as the input of temporal evaluation module. Given a feature sequence F, temporal evaluation module can generate three probability sequences $P_S = \{p^s_{t_n}\}_{n=1}^{l_s}$, $P_E = \{p^e_{t_n}\}_{n=1}^{l_s}$ and $P_A = \{p^a_{t_n}\}_{n=1}^{l_s}$, where $p^s_{t_n}$, $p^e_{t_n}$ and $p^a_{t_n}$ are respectively starting, ending and actionness probabilities in time t_n.
时间卷积层可以简单地表示为$Conv(c_f, c_k, Act)$，其中c_f，c_k和Act分别是过滤数字，内核大小和时间卷积层的激活函数。如图2（b）所示，时间评估模块可以定义为$Conv(512, 3, Relu) \rightarrow Conv(512, 3, Relu) \rightarrow Conv(3, 1, Sigmoid)$，其中三个层具有相同的步幅大小1。在最后一层中使用Sigmoid激活的三个过滤器被用作分类器以分别生成起始，结束和动作概率。为了便于计算，我们将特征序列划分为非重叠窗口作为时间评估模块的输入。给定特征序列F，时间评估模块可以生成三个概率序列$P_S = \{p^s_{t_n}\}_{n=1}^{l_s}$，$P_E = \{p^e_{t_n}\}_{n=1}^{l_s}$和$P_A = \{p^a_{t_n}\}_{n=1}^{l_s}$，其中$p^s_{t_n}$，$p^e_{t_n}$和$p^a_{t_n}$分别是时间$t_n$中的开始，结束和动作概率。

Proposal generation module. The goal of proposal generation module is to generate candidate proposals and construct corresponding proposal-level feature. We achieve this goal in two steps. First we locate temporal locations with high boundary probabilities, and combine these locations to form proposals. Then for each proposal, we construct Boundary-Sensitive Proposal (BSP) feature.
提案生成模块。提议生成模块的目标是生成候选提议并构建相应的提议级功能。我们分两步实现这一目标。首先，我们找到具有高边界概率的时间位置，并将这些位置组合以形成提议。然后，对于每个提议，我们构建边界敏感提议（BSP）功能。

As shown in Fig 3(a), to locate where an action likely to start, for starting probabilities sequence P_S, we record all temporal location t_n where $p^s_{t_n}$ (1) has high score: $p^s_{t_n} > 0.9$ or (2) is a probability peak: $p^s_{t_n} > p^s_{t_{n-1}}$ and $p^s_{t_n} > p^s_{t_{n+1}}$. These locations are grouped into candidate starting locations set $B_S = \{t_{s,i}\}_{i=1}^{N_S}$, where N_S is the number of candidate starting locations. Using same rules, we can generate candidate ending locations set B_E from ending probabilities sequence P_E. Then, we generate temporal regions via combing each starting location t_s from B_S and each ending location t_e from B_S. Any temporal region $\lfloor t_s, t_e \rfloor$ satisfying $d = t_e - t_s \in [d_{min}, d_{max}]$ is denoted as a candidate proposal φ, where d_{min} and d_{max} are minimum and maximum durations of ground truth action instances in dataset. Thus we can get candidate proposals set $\Psi_p = \{\varphi_i\}_{i=1}^{N_p}$, where N_p is the number of proposals.
如图3（a）所示，为了定位可能开始的动作，为了开始概率序列P_S，我们记录$p^s_{t_n}$（1）具有高分的所有时间位置t_n：$p^s_{t_n} > 0.9$或（2）是概率峰值：$p^s_{t_n} > p^s_{t_{n-1}}$和$p^s_{t_n} > p^s_{t_{n+1}}$。这些位置被分组到设置为$B_S = \{t_{s,i}\}_{i=1}^{N_S}$的候选起始位置，其中$N_S$是候选起始位置的数量。使用相同的规则，我们可以从结束概率序列P_E生成设置B_E的候选结束位置。然后，我们通过组合来自B_S的每个起始位置t_s和来自B_S的每个结束位置t_e来生成时间区域。满足$d = t_e - t_s \in [d_{min}, d_{max}]$的任何时间区域$\lfloor t_s, t_e \rfloor$被表示为候选提议φ，其中$d_{min}$和$d_{max}$是数据集中的

地面实况动作实例的最小和最大持续时间。因此，我们可以将候选提案设置为$\Psi_p = \{\varphi_i\}_{i=1}^{N_p}$，其中$N_p$是提案数。

To construct proposal-level feature as shown in Fig 3(b), for a candidate proposal φ, we denote its center region as $r_C = [t_s, t_e]$ and its starting and ending region as $r_S = [t_s - d/5, t_s + d/5]$ and $r_E = [t_e - d/5, t_e + d/5]$ separately. Then, we sample the actionness sequence P_A within r_c as f_c^A by linear interpolation with 16 points. In starting and ending regions, we also sample actionness sequence with 8 linear interpolation points and get f_s^A and f_e^A separately. Concatenating these vectors, we can get Boundary-Sensitive Proposal (BSP) feature $f_{BSP} = (f_s^A, f_c^A, f_e^A)$ of proposal φ. BSP feature is highly compact and contains rich semantic information about corresponding proposal. Then we can represent a proposal as $\varphi = (t_s, t_e, f_{BSP})$.
为了构建投标级特征，如图3（b）中，用于候选提案φ，我们表示其中心区域为$r_C = [t_s, t_e]$和其起始和结束区域作为单独$r_S = [t_s - d/5, t_s + d/5]$和$r_E = [t_e - d/5, t_e + d/5]$。然后，我们通过16点线性插值将$r_c$内的动作序列$P_A$作为$f_c^A$进行采样。在起始区域和结束区域中，我们还采用8个线性插值点对动作序列进行采样，并分别得到f_s^A和f_e^A。连接这些向量，我们可以获得提议φ的边界敏感提议（BSP）特征$f_{BSP} = (f_s^A, f_c^A, f_e^A)$。BSP功能非常紧凑，包含有关相应提案的丰富语义信息。然后我们可以将提案表示为$\varphi = (t_s, t_e, f_{BSP})$。

Proposal evaluation module. The goal of proposal evaluation module is to evaluate the confidence score of each proposal whether it contains an action instance within its duration using BSP feature. We adopt a simple multilayer perceptron model with one hidden layer as shown in Fig 2(b). Hidden layer with 512 units handles the input of BSP feature f_{BSF} with Relu activation. The output layer outputs confidence score p_{conf} with sigmoid activation, which estimates the overlap extent between candidate proposal and ground truth action instances. Thus, a generated proposal can be denoted as $\varphi = (t_s, t_e, p_{conf}, p_{t_s}^s, p_{t_e}^e)$, where $p_{t_s}^s$ and $p_{t_e}^e$ are starting and ending probabilities in t_s and t_e separately. These scores are fused to generate final score during prediction.
提案评估模块。提案评估模块的目标是使用BSP功能评估每个提案在其持续时间内是否包含操作实例的信度得分。我们采用具有一个隐藏层的简单多层感知器模型，如图2（b）所示。具有512个单元的隐藏层处理带有Relu激活的BSP功能f_{BSF}的输入。输出层输出具有sigmoid激活的信息得分p_{conf}，其估计候选提议和地面实况动作实例之间的重叠范围。因此，生成的提议可以表示为$\varphi = (t_s, t_e, p_{conf}, p_{t_s}^s, p_{t_e}^e)$，其中$p_{t_s}^s$和$p_{t_e}^e$分别是$t_s$和$t_e$中的开始和结束概率。融合这些分数以在预测期间产生最终分数。

3.4 Training of BSN
3.4 BSN培训

In BSN, temporal evaluation module is trained to learn local boundary and actionness probabilities from video features simultaneously. Then based on probabilities sequence generated by trained temporal evaluation module, we can generate proposals and corresponding BSP features and train the proposal evaluation module to learn the confidence score of proposals. The training details are introduced in this section.
在BSN中，训练时间评估模块以同时从视频特征学习局部边界和动作概率。然后基于训练后的时态评估模块生成的概率序列，我们可以生成提议和相应的BSP特征，并训练提案评估模块，以了解提案的信心分数。本节介绍了培训详细信息。

Temporal evaluation module. Given a video X, we compose a snippets sequence S with length l_s and extract feature sequence F from it. Then we slide windows with length $l_w = 100$ in feature sequence without overlap. A window is denoted as $\omega = \{F_\omega, \Psi_\omega\}$, where F_ω and Ψ_ω are feature sequence and annotations within the window separately.

For ground truth action instance $\varphi_g = (t_s, t_e)$ in Ψ_ω, we denote its region as action region r_g^a and its starting and ending region as $r_g^s = [t_s - d_g/10, t_s + d_g/10]$ and $r_g^e = [t_e - d_g/10, t_e + d_g/10]$ separately, where $d_g = t_e - t_s$.
时间评估模块。给定视频X，我们组成一个长度为l_s的片段序列S，并从中提取特征序列F.然后我们在特征序列中滑动长度为$l_w = 100$的窗口，没有重叠。窗口表示为$\omega = \{F_\omega, \Psi_\omega\}$，其中$F_\omega$和$\Psi_\omega$分别是窗口内的特征序列和注释。对于$\Psi_\omega$中的地面实况动作实例$\varphi_g = (t_s, t_e)$，我们将其区域表示为动作区域$r_g^a$，将其起始和结束区域分别表示为$r_g^s = [t_s - d_g/10, t_s + d_g/10]$和$r_g^e = [t_e - d_g/10, t_e + d_g/10]$，其中$d_g = t_e - t_s$。

Taking F_ω as input, temporal evaluation module generates probabilities sequence $P_{S,\omega}$, $P_{E,\omega}$ and $P_{A,\omega}$ with same length l_w. For each temporal location t_n within F_ω, we denote its region as $r_{t_n} = [t_n - d_s/2, t_n + d_s/2]$ and get corresponding probability scores $p_{t_n}^s$, $p_{t_n}^e$ and $p_{t_n}^a$ from $P_{S,\omega}$, $P_{E,\omega}$ and $P_{A,\omega}$ separately, where $d_s = t_n - t_{n-1}$ is temporal interval between two snippets. Then for each r_{t_n}, we calculate its IoP ratio with r_g^a, r_g^s and r_g^e of all φ_g in Ψ_ω separately, where IoP is defined as the overlap ratio with groundtruth proportional to the duration of this proposal. Thus we can represent information of t_n as $\phi_n = (p_{t_n}^a, p_{t_n}^s, p_{t_n}^e, g_{t_n}^a, g_{t_n}^s, g_{t_n}^e)$, where $g_{t_n}^a$, $g_{t_n}^s$, $g_{t_n}^e$ are maximum matching overlap IoP of action, starting and ending regions separately.
以F_ω作为输入，时间评估模块生成具有相同长度l_w的概率序列$P_{S,\omega}$，$P_{E,\omega}$和$P_{A,\omega}$。对于F_ω内的每个时间位置t_n，我们将其区域表示为$r_{t_n} = [t_n - d_s/2, t_n + d_s/2]$，并分别从$P_{S,\omega}$，$P_{E,\omega}$和$P_{A,\omega}$获得相应的概率分数$p_{t_n}^s$，$p_{t_n}^e$和$p_{t_n}^a$，其中$d_s = t_n - t_{n-1}$是两个片段之间的时间间隔。然后，对于每个$r_{t_n}$，我们分别使用$\Psi_\omega$中所有$\varphi_g$的$r_g^a$，$r_g^s$和$r_g^e$来计算其IoP比率，其中IoP被定义为与本提案的持续时间成比例的地面重叠率。因此，我们可以将t_n的信息表示为$\phi_n = (p_{t_n}^a, p_{t_n}^s, p_{t_n}^e, g_{t_n}^a, g_{t_n}^s, g_{t_n}^e)$，其中$g_{t_n}^a$，$g_{t_n}^s$，$g_{t_n}^e$最大匹配动作的重叠IoP，分别开始和结束区域。

Given a window of matching information as $\Phi_\omega = \{\phi_n\}_{n=1}^{l_s}$, we can define training objective of this module as a three-task loss function. The overall loss function consists of actionness loss, starting loss and ending loss:
给定一个匹配信息窗口作为$\Phi_\omega = \{\phi_n\}_{n=1}^{l_s}$，我们可以将该模块的训练目标定义为三任务丢失函数。整体损失函数包括行动损失，起始损失和期末损失：

$$L_{TEM} = \lambda \cdot L_{bl}^{action} + L_{bl}^{start} + L_{bl}^{end}, \tag{1}$$

where λ is the weight term and is set to 2 in BSN. We adopt the sum of binary logistic regression loss function L_{bl} for all three tasks, which can be denoted as:
其中λ是权重项，在BSN中设置为2。我们对所有三个任务采用二元逻辑回归损失函数L_{bl}的总和，可以表示为：

$$L_{bl} = \frac{1}{l_w} \sum_{i=1}^{l_w} \left(\alpha^+ \cdot b_i \cdot log(p_i) + \alpha^- \cdot (1 - b_i) \cdot log(1 - p_i) \right), \tag{2}$$

where $b_i = sign(g_i - \theta_{IoP})$ is a two-values function for converting matching score g_i to $\{0, 1\}$ based on threshold θ_{IoP}, which is set to 0.5 in BSN. Let $l^+ = \sum g_i$ and $l^- = l_w - l^+$, we can set $\alpha^+ = \frac{l_w}{l^-}$ and $\alpha^- = \frac{l_w}{l^+}$, which are used for balancing the effect of positive and negative samples during training.

Proposal evaluation module. Using probabilities sequences generated by trained temporal evaluation module, we can generate proposals using proposal generation module: $\Psi_p = \{\varphi_n = (t_s, t_e, f_{BSP})\}_{n=1}^{N_p}$. Taking f_{BSF} as input, for a

proposal φ, confidence score p_{conf} is generated by proposal evaluation module. Then we calculate its Intersection-over-Union (IoU) with all φ_g in Ψ_g, and denote the maximum overlap score as g_{iou}. Thus we can represent proposals set as $\Psi_p = \{\varphi_n = \{t_s, t_e, p_{conf}, g_{iou}\}\}_{n=1}^{N_p}$. We split Ψ_p into two parts based on g_{iou} : Ψ_p^{pos} for $g_{iou} > 0.7$ and Ψ_p^{neg} for $g_{iou} < 0.3$. For data balancing, we take all proposals in Ψ_p^{pos} and randomly sample the proposals in Ψ_p^{neg} to insure the ratio between two sets be nearly 1:2.

提案评估模块。使用由训练的时间评估模块生成的概率序列，我们可以使用提议生成模块生成提议：$\Psi_p = \{\varphi_n = (t_s, t_e, f_{BSP})\}_{n=1}^{N_p}$。以$f_{BSF}$作为输入，对于提议φ，由提议评估模块生成信息得分$p_{conf}$。然后我们用$\Psi_g$中的所有$\varphi_g$计算它的联合交叉（IoU），并将最大重叠分数表示为g_{iou}。因此，我们可以表示设置为$\Psi_p = \{\varphi_n = \{t_s, t_e, p_{conf}, g_{iou}\}\}_{n=1}^{N_p}$的提案。我们将$\Psi_p$分为两部分，基于$g_{iou}$：$\Psi_p^{pos}$用于$g_{iou} > 0.7$，$\Psi_p^{neg}$用于$g_{iou} < 0.3$。对于数据平衡，我们采用$\Psi_p^{pos}$中的所有提案并随机抽样$\Psi_p^{neg}$中的提案，以确保两组之间的比率接近1：2。

The training objective of this module is a simple regression loss, which is used to train a precise confidence score prediction based on IoU overlap. We can define it as:

该模块的训练目标是简单的回归损失，用于根据IoU重叠训练精确的置信分数预测。我们可以将它定义为：

$$L_{PEM} = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} (p_{conf,i} - g_{iou,i})^2, \tag{3}$$

where N_{train} is the number of proposals used for training.

其中N_{train}是用于培训的提案数量。

3.5 Prediction and Post-processing

3.5预测和后处理

During prediction, we use BSN with same procedures described in training to generate proposals set $\Psi_p = \{\varphi_n = (t_s, t_e, p_{conf}, p_{t_s}^s, p_{t_e}^e)\}_{n=1}^{N_p}$, where N_p is the number of proposals. To get final proposals set, we need to make score fusion to get final confidence score, then suppress redundant proposals based on these score.

在预测期间，我们使用BSN与培训中描述的相同程序生成提议集$\Psi_p = \{\varphi_n = (t_s, t_e, p_{conf}, p_{t_s}^s, p_{t_e}^e)\}_{n=1}^{N_p}$，其中$N_p$是提案数。为了获得最终的建议，我们需要进行分数融合以获得最终的置信分数，然后根据这些分数来抑制冗余的提议。

Score fusion for retrieving. To achieve better retrieving performance, for each candidate proposal φ, we fuse its confidence score with its boundary probabilities by multiplication to get the final confidence score p_f:

用于检索的分数融合。为了获得更好的检索性能，对于每个候选提议φ，我们通过乘法将其信度得分与其边界概率融合，以获得最终的信度得分p_f：

$$p_f = p_{conf} \cdot p_{t_s}^s \cdot p_{t_e}^e. \tag{4}$$

After score fusion, we can get generated proposals set $\Psi_p = \{\varphi_n = (t_s, t_e, p_f)\}_{n=1}^{N_p}$, where p_f is used for proposals retrieving. In section 4.2, we explore the recall performance with and without confidence score generated by proposal evaluation module.

在得分融合之后，我们可以获得生成的提议集$\Psi_p = \{\varphi_n = (t_s, t_e, p_f)\}_{n=1}^{N_p}$，其中$p_f$用于提案检索。在4.2节中，我们探讨了提案评估模块生成和不提供评估分数的召回绩效。

Redundant proposals suppression. Around a ground truth action instance, we may generate multiple proposals with different temporal overlap. Thus we need to suppress redundant proposals to obtain higher recall with fewer proposals.
冗余提案压制。围绕一个地面实况行动实例，我们可能会生成具有不同时间重叠的多个提案。因此，我们需要压制多余的提案以获得更高的召回率和更少的提案。

Soft-NMS [33] is a recently proposed non-maximum suppression (NMS) algorithm which suppresses redundant results using a score decaying function. First all proposals are sorted by their scores. Then proposal φ_m with maximum score is used for calculating overlap IoU with other proposals, where scores of highly overlapped proposals is decayed. This step is recursively applied to the remaining proposals to generate rescored proposals set. The Gaussian decaying function of Soft-NMS can be denoted as:
Soft-NMS [33]是最近提出的非最大抑制（NMS）算法，其使用分数衰减函数来抑制冗余结果。首先，所有提案都按其分数排序。然后使用具有最大分数的提议φ_m来计算与其他提议的重叠IoU，其中高度重叠的提议的分数被衰减。将该步骤递归地应用于剩余的提议以生成重新设定的提议集。Soft-NMS的高斯衰减函数可表示为：

$$p'_{f,i} = \begin{cases} p_{f,i}, & iou(\varphi_m, \varphi_i) < \theta \\ p_{f,i} \cdot e^{-\frac{iou(\varphi_m, \varphi_i)^2}{\varepsilon}}, & iou(\varphi_m, \varphi_i) \geq \theta \end{cases} \tag{5}$$

where ε is parameter of Gaussian function and θ is pre-fixed threshold. After suppression, we get the final proposals set $\Psi'_p = \left\{\varphi_n = (t_s, t_e, p'_f)\right\}_{n=1}^{N_p}$.

4 Experiments
4实验

4.1 Dataset and setup
4.1数据集和设置

Dataset. ActivityNet-1.3 [1] is a large dataset for general temporal action proposal generation and detection, which contains 19994 videos with 200 action classes annotated and was used in the ActivityNet Challenge 2016 and 2017. ActivityNet-1.3 is divided into training, validation and testing sets by ratio of 2:1:1. THUMOS14 [2] dataset contains 200 and 213 temporal annotated untrimmed videos with 20 action classes in validation and testing sets separately. The training set of THUMOS14 is the UCF-101 [34], which contains trimmed videos for action recognition task. In this section, we compare our method with state-of-the-art methods on both ActivityNet-1.3 and THUMOS14.
数据集。 ActivityNet-1.3 [1]是一个用于一般时间行动建议生成和检测的大型数据集，其中包含19994个带有200个动作类注释的视频，并在2016年和2017年的ActivityNet挑战中使用。ActivityNet-1.3按比例2：1：1分为训练，验证和测试集。THUMOS14 [2]数据集包含200和213个临时带注释的未修剪视频，分别在验证和测试集中有20个动作类。THUMOS14的训练集是UCF-101 [34]，其中包含用于动作识别任务的修剪视频。在本节中，我们将我们的方法与ActivityNet-1.3和THUMOS14上的最新方法进行比较。

Evaluation metrics. In temporal action proposal generation task, Average Recall (AR) calculated with multiple IoU thresholds is usually used as evaluation metrics. Following conventions, we use IoU thresholds set $|0.5 : 0.05 : 0.95|$ in ActivityNet-1.3 and $|0.5 : 0.05 : 1.0|$ in THUMOS14. To evaluate the relation between recall and proposals number, we evaluate AR with Average Number of proposals (AN) on both datasets, which is denoted as AR@AN. On ActivityNet-1.3, area under the AR vs. AN curve (AUC) is also used as metrics, where AN varies from 0 to 100.

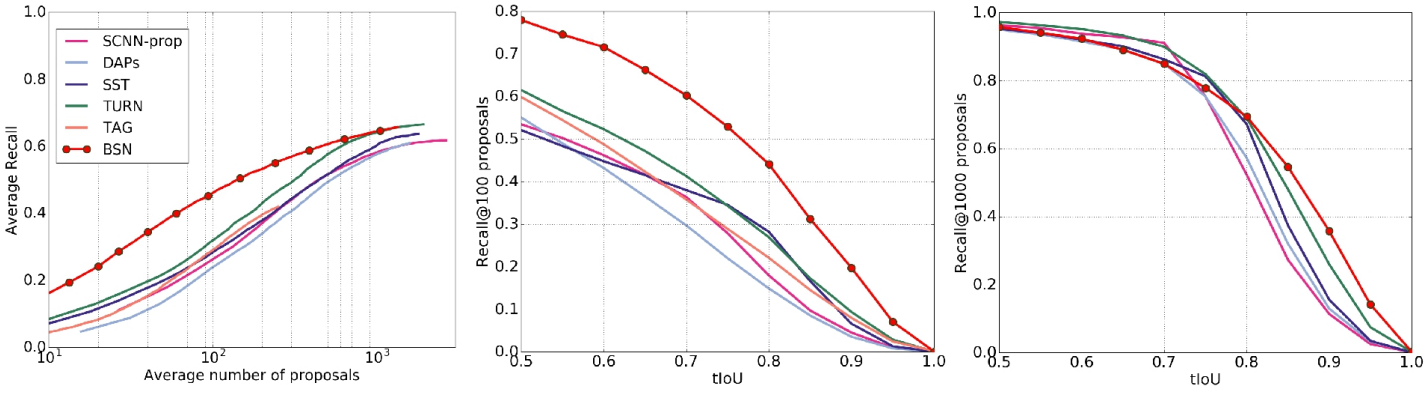
评估指标。在时间动作提议生成任务中，使用多个IoU阈值计算的平均召回（AR）通常用作评估度量。遵循约定，我们在THUMOS14中使用ActivityNet-1.3中的$|0.5 : 0.05 : 0.95|$和THUMOS14中的$|0.5 : 0.05 : 1.0|$设置IoU阈值。为了评估召回和建议编号之间的关系，我们在两个数据集上评估具有平均建议数（AN）的AR，表示为AR @ AN。在ActivityNet-1.3上，AR与AN曲线（AUC）下的面积也用作度量，其中AN在0到100之间变化。

In temporal action detection task, mean Average Precision (mAP) is used as evaluation metric, where Average Precision (AP) is calculated on each action class respectively. On ActivityNet-1.3, mAP with IoU thresholds $\{0.5, 0.75, 0.95\}$ and average mAP with IoU thresholds set $|0.5 : 0.05 : 0.95|$ are used. On THUMOS14, mAP with IoU thresholds $\{0.3, 0.4, 0.5, 0.6, 0.7\}$ is used.

在时间动作检测任务中，平均精度（mAP）用作评估度量，其中分别针对每个动作类计算平均精度（AP）。在ActivityNet-1.3上，使用具有IoU阈值$\{0.5, 0.75, 0.95\}$的mAP和具有设置$|0.5 : 0.05 : 0.95|$的IoU阈值的平均mAP。在THUMOS14上，使用具有IoU阈值$\{0.3, 0.4, 0.5, 0.6, 0.7\}$的mAP。

Implementation details. For visual feature encoding, we use the two-stream network Fig. 4: Comparison of our proposal generation method with other state-of-the-art methods in THUMOS14 dataset. (left) BSN can achieve significant performance gains with relatively few proposals. (center) Recall with 100 proposals vs tIoU figure shows that with few proposals, BSN gets performance improvements in both low and high tIoU. (right) Recall with 1000 proposals vs tIoU figure shows that with large number of proposals, BSN achieves improvements mainly while tIoU > 0.8.

实施细节。对于视觉特征编码，我们使用双流网络图4：我们的提议生成方法与THUMOS14数据集中其他最先进方法的比较。 （左）BSN可以通过相对较少的提议获得显着的性能提升。 （中）回顾100个提案与tIoU图表显示，由于提案很少，BSN在低和高的tIUU都得到了性能提升。 （右）召回1000项提案与tIoU图表显示，由于提案数量众多，BSN主要在tIoU> 0.8时实现了改进。



On ActivityNet-1.3, since the duration of videos are limited, we follow [39] to rescale the feature sequence of each video to new length $l_w = 100$ by linear interpolation, and the duration of corresponding annotations to range [0,1]. In BSN,

temporal evaluation module and proposal evaluation module are both implemented using Tensorflow [40]. On both datasets, temporal evaluation module is trained with batch size 16 and learning rate 0.001 for 10 epochs, then 0.0001 for another 10 epochs, and proposal evaluation module is trained with batch size 256 and same learning rate. For Soft-NMS, we set the threshold θ to 0.8 on ActivityNet-1.3 and 0.65 on THUMOS14 by empirical validation, while ε in Gaussian function is set to 0.75 on both datasets.
在ActivityNet-1.3上，由于视频的持续时间有限，我们按照[39]通过线性插值将每个视频的特征序列重新缩放到新长度$l_w = 100$，并将相应注释的持续时间重新调整到范围[0,1]。在BSN中，时态评估模块和提议评估模块都是使用Tensor流程[40]实现的。在两个数据集上，时间评估模块以批量大小16和学习率0.001训练10个时期，然后0.0001训练另外10个时期，并且提议评估模块以批量大小256和相同学习率训练。对于Soft-NMS，我们通过经验验证将ActivityNet-1.3上的阈值θ设置为0.8，将THUMOS14上的阈值设置为0.65，而在两个数据集上将高斯函数中的ε设置为0.75。

4.2 Temporal Proposal Generation
4.2时间提案生成

Taking a video as input, proposal generation method aims to generate temporal proposals where action instances likely to occur. In this section, we compare our method with state-of-the-art methods and make external experiments to verify effectiveness of BSN. Comparison with state-of-the-art methods. As aforementioned, a good proposal generation method should generate and retrieve proposals to cover ground truth action instances with high recall and high temporal overlap using relatively few proposals. We evaluate these methods in two aspects.
将视频作为输入，提议生成方法旨在生成可能发生动作实例的时间提议。在本节中，我们将我们的方法与最先进的方法进行比较，并进行外部实验以验证BSN的有效性。与最先进的方法进行比较。如上所述，良好的提案生成方法应该使用相对较少的提议生成和检索提议以涵盖具有高召回率和高时间重叠的地面实况动作实例。我们从两个方面评估这些方法。

First we evaluate the ability of our method to generate and retrieve proposals with high recall, which is measured by average recall with different number of proposals (AR@AN) and area under AR-AN curve (AUC). We list the comparison results of ActivityNet-1.3 and THUMOS14 in Table 1 and Table 2 respectively, and plot the average recall against average number of proposals curve of THUMOS14 in Fig 4 (left). On THUMOS14, our method outperforms other state-of-the-art proposal methods when proposal number varies from 10 to 1000. Especially, when average number of proposals is 50, our method significantly improves average recall from 21.86% to 37.46% by 15.60%. On ActivityNet-1.3, our method outperforms other state-of-the-art proposal generation methods on both validation and testing set.
首先，我们评估我们的方法生成和检索具有高召回率的提议的能力，其通过具有不同提议数量（AR @ AN）和AR-AN曲线下面积（AUC）的平均召回来测量。我们分别列出了表1和表2中ActivityNet-1.3和THUMOS14的比较结果，并绘制了图4（左）中THUMOS14的平均召回率与平均提议数曲线的关系曲线。在THUMOS14上，当提案编号从10变为1000时，我们的方法优于其他最先进的提案方法。特别是，当平均提案数为50时，我们的方法显着提高了平均召回率21.86％至37.46％，提高了15.60％。在ActivityNet-1.3上，我们的方法在验证和测试集上都优于其他最先进的提议生成方法。

Second, we evaluate the ability of our method to generate and retrieve proposals with high temporal overlap, which is measured by recall of multiple IoU thresholds. We plot the recall against IoU thresholds curve with 100 and 1000

proposals in Fig 4 (center) and (right) separately. Fig 4 (center) suggests that our method achieves significant higher recall than other methods with 100 proposals when IoU threshold varied from 0.5 to 1.0. And Fig 4 (right) suggests that with 1000 proposals, our method obtains the largest recall improvements when IoU threshold is higher than 0.8.
其次，我们评估了我们的方法生成和检索具有高时间重叠的提议的能力，其通过调用多个IoU阈值来测量。我们将IoU阈值曲线的召回与图4（中）和（右）中的100和1000个提案分别绘制。图4（中）表明，当IoU阈值从0.5变为1.0时，我们的方法比100个提案的其他方法实现了更高的召回率。图4（右）表明，当IoU阈值高于0.8时，我们的方法有1000个提案，获得了最大的召回改进。

Table 1: Comparison between our method with other state-of-the-art proposal generation methods on validation set of ActivityNet-1.3 in terms of AR@AN and AUC.

Method	Zhao et al. [24]	Dai et al. [42]	Yao et al. [43]	Lin et al. [39]	BSN
AR@100 (val)	63.52	-	-	73.01	**74.16**
AUC (val)	53.02	59.58	63.12	64.40	**66.17**
AUC (test)	-	61.56	64.18	64.80	**66.26**

Table 2: Comparison between our method with other state-of-the-art proposal generation methods on THUMOS14 in terms of AR@AN.

Feature	Method	@50	@100	@200	@500	@1000
C3D	DAPs [5]	13.56	23.83	33.96	49.29	57.64
C3D	SCNN-prop [7]	17.22	26.17	37.01	51.57	58.20
C3D	SST [3]	19.90	28.36	37.90	51.58	60.27
C3D	TURN [6]	19.63	27.96	38.34	53.52	**60.75**
C3D	BSN + Greedy-NMS	27.19	35.38	43.61	53.77	59.50
C3D	BSN + Soft-NMS	**29.58**	**37.38**	**45.55**	**54.67**	59.48
2-Stream	TAG [24]	18.55	29.00	39.61	-	-
Flow	TURN [6]	21.86	31.89	43.02	57.63	64.17
2-Stream	BSN + Greedy-NMS	35.41	43.55	52.23	**61.35**	**65.10**
2-Stream	BSN + Soft-NMS	**37.46**	**46.06**	**53.21**	60.64	64.52

Furthermore, we make some controlled experiments to confirm the contribution of BSN itself in Table 2. For video feature encoding, except for two-stream network, C3D network [12] is also adopted in some works [3,5,6,7]. For NMS method, most previous work adopt Greedy-NMS [41] for redundant proposals suppression. Thus, for fair comparison, we train BSN with feature extracted by C3D network [12] pre-trained on UCF-101 dataset, then perform Greedy-NMS and Soft-NMS on C3D-BSN and original 2Stream-BSN respectively. Results in Table 2 show that (1) C3D-BSN still outperforms other C3D-based methods especially with small proposals number, (2) Soft-NMS only brings small performance promotion than Greedy-NMS, while Greedy-NMS also works well with BSN. These results suggest that the architecture of BSN itself is the main reason for performance promotion rather than input feature and NMS method.
此外，我们进行了一些对照实验，以确定表2中BSN本身的贡献。对于视频特征编码，除了双流网络，C3D网络[12]也被采用在一些作品[3,5,6,7]中。对于NMS方法，大多数先前的工作采用Greedy-NMS [41]来抑制冗余建议。因此，为了公平比较，我们训练BSN具有由在NCF-101数据集上预训练的C3D网络[12]提取的特征，然后分别在C3D-BSN和原始2Stream-BSN上执行Greedy-NMS和Soft-NMS。表2中的结果表明（1）C3D-BSN仍然优于其他基

于C3D的方法，特别是提议数量较少，（2）Soft-NMS仅比Greedy-NMS带来小的性能提升，而Greedy-NMS也适用于BSN 。这些结果表明，BSN本身的体系结构是性能提升的主要原因，而不是输入特征和NMS方法。

These results suggest the effectiveness of BSN. And BSN achieves the salient performance since it can generate proposals with (1) flexible temporal duration to cover ground truth action instances with various durations; (2) precise temporal boundary via learning starting and ending probability using temporal convolutional network, which brings high overlap between generated proposals and ground truth action instances; (3) reliable confidence score using BSP feature, which retrieves proposals properly so that high recall and high overlap can be achieved using relatively few proposals. Qualitative examples on THUMOS14 and ActivityNet-1.3 datasets are shown in Fig 5.
这些结果表明了BSN的有效性。并且BSN实现了显着的性能，因为它可以生成具有（1）灵活的时间持续时间的提议，以覆盖具有不同持续时间的地面实况动作实例；（2）通过使用时间卷积网络学习开始和结束概率的精确时间边界，这在生成的提议和地面实况动作实例之间带来高度重叠；（3）使用BSP功能的可靠信任评分，它可以正确地检索建议，以便使用相对较少的建议可以实现高召回率和高重叠率。THUMOS14和ActivityNet-1.3数据集的定性示例如图5所示。

Generalizability of proposals. Another key property of a proposal generation method is the ability to generate proposals for unseen action classes. To evaluate this property, we choose two semantically different action subsets on ActivityNet-1.3: “Sports, Ex ercise, and Recreation” and “Socializing, Relaxing, and Leisure” as seen and unseen subsets separately. Seen subset contains 87 action classes with 4455 training and 2198 validation videos, and unseen subset contains 38 action classes with 1903 training and 896 validation videos. To guarantee the experiment effectiveness, instead of two-stream network, here we adopt C3D network [44] trained on Sports-1M dataset [45] for video features encoding. Using C3D feature, we train BSN with seen and seen+unseen videos on training set separately, then evaluate both models on seen and unseen validation videos separately. As shown in Table 3, there is only slight performance drop in unseen classes, which demonstrates that BSN has great generalizability and can learn a generic concept of temporal action proposal even in semantically different unseen actions.
提案的普遍性。提案生成方法的另一个关键属性是能够为看不见的操作类生成提议。为了评估这个属性，我们在ActivityNet-1.3上选择了两个语义不同的动作子集：“体育，锻炼和娱乐”和“社交，放松和休闲”，分别是看到和看不见的子集。所见的子集包含87个动作类，4455个训练和2198个验证视频，而看不见的子集包含38个动作类，1903个训练和896个验证视频。为了保证实验的有效性，我们采用在Sports-1M数据集[45]上训练的C3D网络[44]代替双流网络进行视频特征编码。使用C3D功能，我们分别在训练集上训练BSN上看到和看过的+看不见的视频，然后分别在看到和看不见的验证视频上评估两个模型。如表3所示，在看不见的类中只有轻微的性能下降，这表明BSN具有很好的普遍性，并且即使在语义上不同的看不见的动作中也可以学习时间动作提议的一般概念。

Table 3: Generalization evaluation of BSN on ActivityNet-1.3. *Seen* subset: “Sports, Exercise, and Recreation”; *Unseen* subset: “Socializing, Relaxing, and Leisure”.

	Seen (validation)		*Unseen* (validation)	
	AR@100	AUC	AR@100	AUC
BSN trained with *Seen* + *Unseen* (training)	72.40	63.80	71.84	63.99
BSN trained with *Seen* (training)	72.42	64.02	**71.32**	**63.38**

Effectiveness of modules in BSN. To evaluate the effectiveness of temporal evaluation module (TEM) and proposal

evaluation module (PEM) in BSN, we demonstrate experiment results of BSN with and without PEM in Table 4, where TEM is used in both results. These results show that: (1) using only TEM without PEM, BSN can also reach considerable recall performance over state-of-the-art methods; (2) PEM can bring considerable further performance promotion in BSN. These observations suggest that TEM and PEM are both effective and indispensable in BSN.
BSN中模块的有效性。为了评估时间评估模块（TEM）和建议评估模块（PEM）在BSN中的有效性，我们在表4中展示了具有和不具有PEM的BSN的实验结果，其中在两个结果中使用TEM。这些结果表明：（1）仅使用没有PEM的TEM，BSN也可以通过最先进的方法获得相当大的召回性能; （2）PEM可以在BSN中带来可观的进一步表现。这些观察结果表明，TEM和PEM在BSN中都是有效且必不可少的。

Boundary-Sensitive Proposal feature. BSP feature is used in proposal evaluation module to evaluate the confidence scores of proposals. In Table 4, we also make ablation studies of the contribution of each component in BSP. These results suggest that although BSP feature constructed from boundary regions contributes less improvements than center region, best recall performance is achieved while PEM is trained with BSP constructed from both boundary and center region.
边界敏感提案功能。 BSP功能用于提案评估模块，用于评估提案的置信度分数。在表4中，我们还对BSP中每种成分的贡献进行了消融研究。这些结果表明，尽管从边界区域构建的BSP特征比中心区域贡献得更少，但是在用边界和中心区域构建的BSP训练PEM的同时实现了最佳的回忆性能。

4.3 Action Detection with Our Proposals
4.3使用我们的建议检测行动

To further evaluate the quality of proposals generated by BSN, we put BSN proposals into “detection by classifying proposals” temporal action detection framework with state-of-the-art action classifier, where temporal boundaries of detection results are provided by our proposals. On ActivityNet-1.3, we use top-2 video-level class generated by classification model [46]1 for all proposals in a video. On THUMOS14, we use top-2 video-level classes generated by UntrimmedNet [48] for proposals generated by BSN and other methods.Following previous works, on THUMOS14, we also implement SCNN-classifier on BSN proposals for proposal-level classification and adopt Greedy NMS as [7]. We use 100 and 200 proposals per video on ActivityNet-1.3 and THUMOS14 datasets separately.
为了进一步评估BSN生成的提案的质量，我们将BSN提案置于“通过分类提案检测”时间动作检测框架中，采用最先进的动作分类器，其中检测结果的时间边界由我们的提议提供。在ActivityNet-1.3上，我们使用分类模型[46] 1生成的前2个视频级别类来显示视频中的所有提案。在THUMOS14上，我们使用由UntrimmedNet [48]生成的前2个视频级别类别来生成BSN和其他方法生成的提案。继THUMOS14之前的工作之后，我们还在BSN提案上实施SCNN分类，用于提案级别的分类并采用贪心NMS [7]。我们分别在ActivityNet-1.3和THUMOS14数据集上为每个视频使用100和200个提案。

1 Previously, we adopted classification results from result files of [47]. Recently we found that the classification accuracy of these results are unexpected high. Thus we replace it with classification results of [46] and updated all related experiments accordingly.
1之前，我们采用了[47]的结果文件的分类结果。最近我们发现这些结果的分类准确度出乎意料的高。因此，我们用[46]的分类结果代替它，并相应地更新所有相关实验。

Table 4: Study of effectiveness of modules in BSN and contribution of components in BSP feature on THUMOS14, where PEM is trained with BSP feature constructed by Boundary region (f As , f Ae) and Center region (f_c^A) independently and jointly.

表4：在THUMOS14中研究BSN中模块的有效性和BSP特征中组件的贡献，其中PEM由边界区域（f As，f Ae）和中心区域（f_c^A）独立地和联合地构建的BSP特征训练。

	Boundary	*Center*	@50	@100	@200	@500	@1000
BSN without PEM			30.72	40.52	48.63	57.78	63.04
BSN with PEM	✓		35.61	44.86	52.46	60.00	64.17
		✓	36.80	45.65	52.63	60.18	64.22
	✓	✓	**37.46**	**46.06**	**53.21**	**60.64**	**64.52**

Table 5: Action detection results on validation and testing set of ActivityNet-1.3 in terms of mAP@tIoU and average mAP, where our proposals are combined with videolevel classification results generated by [46].

表5：根据mAP @ tIoU和平均mAP对ActivityNet-1.3的验证和测试集的动作检测结果，其中我们的提议与[46]生成的视频级别分类结果相结合。

	validation				testing
Method	0.5	0.75	0.95	Average	Average
Wang et al. [47]	42.28	3.76	0.05	14.85	14.62
SCC [49]	40.00	17.90	4.70	21.70	19.30
CDC [50]	43.83	25.88	0.21	22.77	22.90
TCN [42]	-	-	-	-	23.58
SSN [51]	39.12	23.48	5.49	23.98	28.28
Lin et al. [39]	44.39	29.65	7.09	29.17	32.26
BSN + [46]	**46.45**	**29.96**	**8.02**	**30.03**	**32.84**

Table 6: Action detection results on testing set of THUMOS14 in terms of mAP@tIoU , where classification results generated by UntrimmedNet [48] and SCNN-classifier [7] are combined with proposals generated by BSN and other methods.

表6：根据mAP @ tIoU对THUMOS14的测试集进行的动作检测结果，其中UntrimmedNet [48]和SCNN-分类器[7]生成的分类结果与BSN和其他方法生成的提议相结合。

	Action Detection Methods				
Detection Method	0.7	0.6	0.5	0.4	0.3
SCNN [7]	5.3	10.3	19.0	28.7	36.3
SMS [30]	-	-	17.8	27.8	36.5
CDC [50]	8.8	14.3	24.7	30.7	41.3
SSAD [25]	7.7	15.3	24.6	35.0	43.0
TCN [42]	9.0	15.9	25.6	33.3	-
R-C3D [52]	9.3	19.1	28.9	35.6	44.8
SS-TAD [26]	9.6	-	29.2	-	45.7
SSN [51]	-	-	29.1	40.8	50.6
CBR [32]	9.9	19.1	31.0	41.3	50.1

Proposal Generation Methods + Action Classifier						
Proposal method	Classifier	0.7	0.6	0.5	0.4	0.3
SST [3]	SCNN-cls	-	-	23.0	-	-
TURN [6]	SCNN-cls	7.7	14.6	25.6	33.2	44.1
SST [3]	UNet	4.7	10.9	20.0	31.5	41.2
TURN [6]	UNet	6.3	14.1	24.5	35.3	46.3
BSN	SCNN-cls	15.0	22.4	29.4	36.6	43.1
BSN	UNet	**20.0**	**28.4**	**36.9**	**45.0**	**53.5**

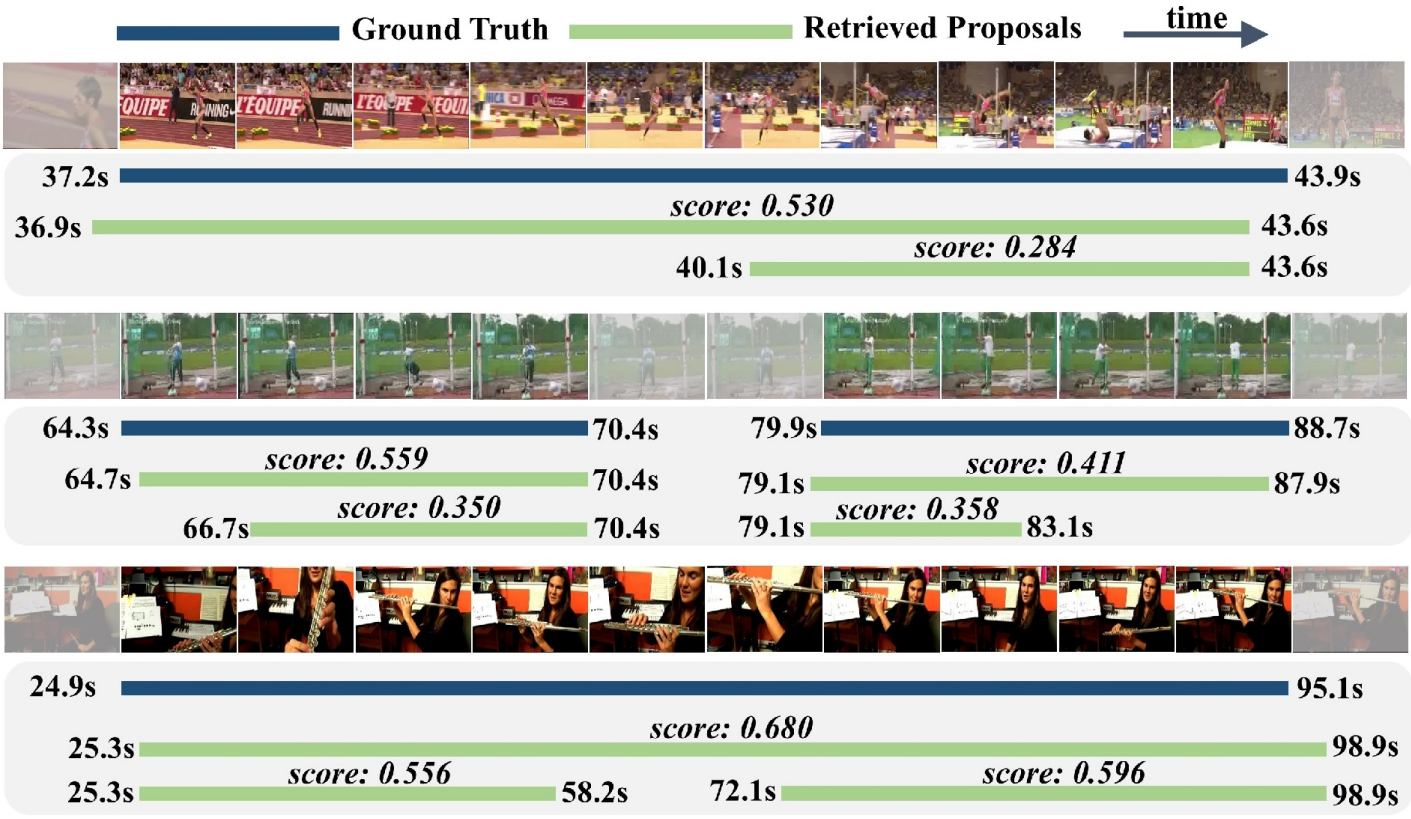


Fig. 5: Qualitative examples of proposals generated by BSN on THUMOS14 (top and middle) and ActivityNet-1.3 (bottom), where proposals are retrieved using postprocessed confidence score.
图5：BSN在THUMOS14（顶部和中部）和ActivityNet-1.3（底部）上生成的提案的定性示例，其中使用后处理的置信分数检索提案。

The comparison results of ActivityNet-1.3 shown in Table 5 suggest that detection framework based on our proposals outperforms other state-of-the-art methods. The comparison results of THUMOS14 shown in Table 6 suggest that (1) using same action classifier, our method achieves significantly better performance than other proposal generation methods; (2) comparing with proposal-level classifier [7], video-level classifier [48] achieves better performance on BSN proposals and worse performance on [3] and [6] proposals, which indicates that confidence scores generated by BSN are more reliable than scores generated by proposal-level classifier, and are reliable enough for retrieving detection results in action detection task; (3) detection framework based on our proposals significantly outperforms state-of-the-art action detection methods, especially when the overlap threshold is high. These results confirm that proposals generated by BSN have high quality and work generally well in detection frameworks.

表5中显示的ActivityNet-1.3的比较结果表明，基于我们提案的检测框架优于其他最先进的方法。表6中显示的THUMOS14的比较结果表明（1）使用相同的动作分类器，我们的方法比其他提议生成方法实现了显着更好的性能; （2）与提案级分类[7]相比，视频级分类[48]在BSN提案上表现更好，在[3]和[6]提案上表现更差，这表明BSN产生的信心分数更可靠比提案级分类器生成的分数，并且足够可靠，可以在动作检测任务中检索检测结果; （3）基于我们的提议的检测框架显着优于最先进的动作检测方法，特别是当重叠阈值很高时。这些结果证实BSN产生的提议具有高质量并且在检测框架中通常很好地工作。

5 Conclusion
5结论

In this paper, we have introduced the Boundary-Sensitive Network (BSN) for temporal action proposal generation. Our method can generate proposals with flexible durations and precise boundaries via directly combing locations with high boundary probabilities, and make accurate retrieving via evaluating proposal confidence score with proposal-level features. Thus BSN can achieve high recall and high temporal overlap with relatively few proposals. In experiments, we demonstrate that BSN significantly outperforms other state-of-the-art proposal generation methods on both THUMOS14 and ActivityNet-1.3 datasets. And BSN can significantly improve the detection performance when used as the proposal stage of a full detection framework.

在本文中，我们引入了边界敏感网络（BSN）来生成时间动作提议。我们的方法可以通过直接组合具有高边界概率的位置来生成具有灵活持续时间和精确边界的提议，并通过使用提议级别特征评估提议置信度得分来进行准确检索。因此，BSN可以用相对较少的提议实现高召回率和高时间重叠。在实验中，我们证明了BSN在THUMOS14和ActivityNet-1.3数据集上都显着优于其他最先进的提议生成方法。当用作完整检测框架的提议阶段时，BSN可以显着提高检测性能。

References
参考

1. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 961–970

1. Caba Heilbron，F.，Escorcia，V.，Ghanem，B.，Carlos Niebles，J。：Activitynet：一个人类活动理解的大型视频基准。在：IEEE计算机视觉和模式识别会议论文集。（2015）961-970

2. Jiang, Y.G., Liu, J., Zamir, A.R., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: Thumos challenge: Action recognition with a large number of classes. In: ECCV Workshop. (2014)

2. Jiang，Y.G.，Liu，J.，Zamir，A.R.，Toderici，G.，Laptev，I.，Shah，M.，Sukthankar，R。：Thumos challenge：具有大量课程的动作识别。在：ECCV研讨会。（2014）

3. Buch, S., Escorcia, V., Shen, C., Ghanem, B., Niebles, J.C.: SST: Single-stream temporal action proposals. In: IEEE International Conference on Computer Vision. (2017)

3. Buch，S.，Escorcia，V.，Shen，C.，Ghanem，B.，Niebles，J.C。：SST：单流时间行动提议。在：IEEE计算机视觉国际会议。（2017）

4. Caba Heilbron, F., Carlos Niebles, J., Ghanem, B.: Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1914–1923
4. Caba Heilbron，F.，Carlos Niebles，J.，Ghanem，B。：快速时间活动建议，用于有效检测未修剪视频中的人类行为。在：IEEE计算机视觉和模式识别会议论文集。（2016）1914-1923

5. Escorcia, V., Heilbron, F.C., Niebles, J.C., Ghanem, B.: Daps: Deep action proposals for action understanding. In: European Conference on Computer Vision, Springer (2016) 768– 784
5. Escorcia，V.，Heilbron，F.C.，Niebles，J.C.，Ghanem，B。：Daps：行动理解的深层行动建议。在：欧洲计算机视觉会议上，Springer（2016）768- 784

6. Gao, J., Yang, Z., Sun, C., Chen, K., Nevatia, R.: Turn tap: Temporal unit regression network for temporal action proposals. arXiv preprint arXiv:1703.06189 (2017)
6. Gao，J.，Yang，Z.，Sun，C.，Chen，K.，Nevatia，R。：Turn tap：Temporal unit regression network for temporal action proposal。 arXiv preprint arXiv：1703.06189（2017）

7. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1049–1058
7. Shou，Z.，Wang，D.，Chang，S.F。：通过多阶段cnns在未修剪的视频中进行时间动作定位。在：IEEE计算机视觉和模式识别会议论文集。（2016）1049-1058

8. Wang, H., Kl¨aser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 3169–3176
8. Wang，H.，Kl¨aser，A.，Schmid，C.，Liu，C.L。：密集轨迹的动作识别。在：计算机视觉和模式识别（CVPR），2011 IEEE Conference on，IEEE（2011）3169-3176

9. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 3551–3558
9. Wang，H.，Schmid，C。：改进轨迹的动作识别。In：IEEE国际计算机视觉会议论文集。（2013）3551-3558

10. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1933–1941
10. Feichtenhofer，C.，Pinz，A.，Zisserman，A。：用于视频动作识别的卷积双流网络融合。在：IEEE计算机视觉和模式识别会议论文集。（2016）1933-1941

11. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems. (2014) 568–576
11. Simonyan，K.，Zisserman，A。：用于视频中动作识别的双流卷积网络。在：神经信息处理系统的进展。（2014）568-576

12. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 4489–4497
12. Tran，D.，Bourdev，L.，Fergus，R.，Torresani，L.，Paluri，M。：用3d卷积网络学习时空特征。In：IEEE国际计算机视觉会议论文集。（2015）4489-4497

13. Wang, L., Xiong, Y., Wang, Z., Qiao, Y.: Towards good practices for very deep two-stream convnets. arXiv preprint arXiv:1507.02159 (2015)
13. Wang，L.，Xiong，Y.，Wang，Z.，Qiao，Y。：建立非常深的双流流行的良好做法。 arXiv preprint arXiv：1507.02159（2015）

14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 580–587
14. Girshick，R.，Donahue，J.，Darrell，T.，Malik，J。：用于精确对象检测和语义分割的丰富特征层次结构。在：IEEE计算机视觉和模式识别会议论文集。 （2014）580-587

15. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1440–1448
15. Girshick，R。：快速r-cnn。 In：IEEE国际计算机视觉会议论文集。 （2015）1440-1448

16. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. (2015) 91–99
16. Ren，S.，He，K.，Girshick，R.，Sun，J。：更快的r-cnn：用区域提案网络进行实时物体检测。在：神经信息处理系统的进展。 （2015）91-99

17. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence 32(9) (2010) 1627–1645
17. Felzenszwalb，P.F.，Girshick，R.B.，McAllester，D.，Ramanan，D。：使用有区别训练的部分模型进行物体检测。关于模式分析和机器智能的IEEE交易32（9）（2010）1627-1645

18. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision 104(2) (2013) 154–171
18. Uijlings，J.R.，van de Sande，K.E.，Gevers，T.，Smeulders，A.W。：选择性搜索物体识别。国际计算机视觉杂志104（2）（2013）154-171

19. Zitnick, C.L., Doll´ar, P.: Edge boxes: Locating object proposals from edges. In: European Conference on Computer Vision, Springer (2014) 391–405
19. Zitnick，C.L.，Doll'ar，P。：边框：从边缘定位对象建议。在：欧洲计算机视觉会议上，Springer（2014）391-405

20. Kuo, W., Hariharan, B., Malik, J.: Deepbox: Learning objectness with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2479–2487
20. Kuo，W.，Hariharan，B.，Malik，J。：Deepbox：用卷积网络学习对象。In：IEEE国际计算机视觉会议论文集。 （2015）2479-2487

21. Lin, T.Y., Doll´ar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. arXiv preprint arXiv:1612.03144 (2016)
21. Lin，T.Y.，Doll'ar，P.，Girshick，R.，He，K.，Hariharan，B.，Belongie，S。：用于物体检测的特征金字塔网络。 arXiv preprint arXiv：1612.03144（2016）

22. Gidaris, S., Komodakis, N.: Locnet: Improving localization accuracy for object detection. In: Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition. (2016) 789–798
22. Gidaris，S.，Komodakis，N。：Locnet：提高物体检测的定位精度。在：IEEE计算机视觉和模式识别会议论文集。（2016）789-798

23. Singh, G., Cuzzolin, F.: Untrimmed video classification for activity detection: submission to activitynet challenge. arXiv preprint arXiv:1607.01979 (2016)
23. Singh，G.，Cuzzolin，F。：用于活动检测的未修剪视频分类：提交到activitynet挑战。 arXiv preprint arXiv：1607.01979（2016）

24. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Lin, D., Tang, X.: Temporal action detection with structured segment networks. arXiv preprint arXiv:1704.06228 (2017)
24. Zhao，Y.，Xiong，Y.，Wang，L.，Wu，Z.，Lin，D.，Tang，X。：结构化分段网络的时间动作检测。 arXiv preprint arXiv：1704.06228（2017）

25. Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: Proceedings of the 25nd ACM international conference on Multimedia. (2017)
25. Lin，T.，Zhao，X.，Shou，Z。：单发时间动作检测。在：第25届ACM国际多媒体会议论文集。（2017）

26. Buch, S., Escorcia, V., Ghanem, B., Fei-Fei, L., Niebles, J.C.: End-to-end, single-stream temporal action detection in untrimmed videos. In: Proceedings of the British Machine Vision Conference. (2017)
26. Buch，S.，Escorcia，V.，Ghanem，B.，Fei-Fei，L.，Niebles，J.C。：未修剪视频中的端到端，单流时间动作检测。在：英国机器视觉会议论文集。（2017）

27. Karaman, S., Seidenari, L., Del Bimbo, A.: Fast saliency based pooling of fisher encoded dense trajectories. In: ECCV THUMOS Workshop. (2014)
27. Karaman，S.，Seidenari，L.，Del Bimbo，A。：基于快速显着性的编码密集轨迹的汇集。在：ECCV THUMOS研讨会。（2014）

28. Oneata, D., Verbeek, J., Schmid, C.: The lear submission at thumos 2014. ECCV THUMOS Workshop (2014)
28. Oneata，D.，Verbeek，J.，Schmid，C。：2014年thumos的学术提交.ECCV THUMOS研讨会（2014）

29. Wang, L., Qiao, Y., Tang, X.: Action recognition and detection by combining motion and appearance features. THUMOS14 Action Recognition Challenge 1 (2014) 2
29. Wang，L.，Qiao，Y.，Tang，X。：结合运动和外观特征的动作识别和检测。 THUMOS14行动认可挑战1（2014）2

30. Yuan, Z., Stroud, J.C., Lu, T., Deng, J.: Temporal action localization by structured maximal sums. arXiv preprint arXiv:1704.04671 (2017)
30. Yuan，Z.，Stroud，J.C.，Lu，T.，Deng，J。：结构化最大和的时间动作定位。 arXiv preprint arXiv：1704.04671（2017）

31. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: towards good practices for deep action recognition. In: European Conference on Computer Vision, Springer (2016) 20–36
31. Wang，L.，Xiong，Y.，Wang，Z.，Qiao，Y.，Lin，D.，Tang，X.，Van Gool，L。：Temporal segment networks：迈向深度行动识别的良好实践。在：欧洲计算机视觉会议，Springer（2016）20-36

32. Gao, J., Yang, Z., Nevatia, R.: Cascaded boundary regression for temporal action detection. arXiv preprint arXiv:1705.01180 (2017)
32. Gao，J.，Yang，Z.，Nevatia，R。：用于时间动作检测的级联边界回归。 arXiv preprint arXiv：1705.01180（2017）

33. Bodla, N., Singh, B., Chellappa, R., Davis, .L.S.: Improving object detection with one line of code. arXiv preprint arXiv:1704.04503 (2017)
33. Bodla，N.，Singh，B.，Chellappa，R.，Davis，.L.S。：用一行代码改进对象检测。 arXiv preprint arXiv：1704.04503（2017）

34. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
34. Soomro，K.，Zamir，A.R.，Shah，M。：Ucf101：来自野外视频的101个人类行动课程的数据集。 arXiv preprint arXiv：1212.0402（2012）

35. Xiong, Y., Wang, L., Wang, Z., Zhang, B., Song, H., Li, W., Lin, D., Qiao, Y., Gool, L.V., Tang, X.: Cuhk & ethz & siat submission to activitynet challenge 2016. arXiv preprint arXiv:1608.00797 (2016)
35. Xiong，Y.，Wang，L.，Wang，Z.，Zhang，B.，Song，H.，Li，W.，Lin，D.，Qiao，Y.，Gool，LV，Tang，X。：Cuhk＆ethz＆siat提交给activitynet challenge 2016. arXiv preprint arXiv：1608.00797（2016）

36. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. (2015) 448–456
36. Ioffe，S.，Szegedy，C。：批量标准化：通过减少内部协变量变化来加速深度网络训练。在：机器学习国际会议。 （2015）448-456

37. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
37. He，K.，Zhang，X.，Ren，S.，Sun，J。：用于图像识别的深度残差学习。在：IEEE计算机视觉和模式识别会议论文集。 （2016）770-778

38. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia, ACM (2014) 675–678
38. Jia，Y.，Shelhamer，E.，Donahue，J.，Karayev，S.，Long，J.，Girshick，R.，Guadarrama，S.，Darrell，T。：Caffe：用于快速特征嵌入的卷积结构。在：第22届ACM国际多媒体会议论文集，ACM（2014）675-678

39. Lin, T., Zhao, X., Shou, Z.: Temporal convolution based action proposal: Submission to activitynet 2017. arXiv preprint arXiv:1707.06750 (2017)
39. Lin，T.，Zhao，X.，Shou，Z。：基于时间卷积的行动建议：2017年活动网的提交.arXiv preprint arXiv：1707.06750（2017）

40. Abadi, M., Agarwal, A., Barham, P., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)
40. Abadi，M.，Agarwal，A.，Barham，P.，et al。：Tensor fl ow：异构分布式系统上的大规模机器学习。 arXiv

preprint arXiv：1603.04467（2016）

41. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition. (2005) 886–893
41. Dalal，N.，Triggs，B。：用于人体检测的定向梯度的直方图。在：IEEE计算机视觉和模式识别会议。（2005）886-893

42. Dai, X., Singh, B., Zhang, G., Davis, L.S., Chen, Y.Q.: Temporal context network for activity localization in videos. In: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE (2017) 5727–5736
42. Dai，X.，Singh，B.，Zhang，G.，Davis，L.S.，Chen，Y.Q。：视频中活动本地化的时间背景网络。在：2017 IEEE计算机视觉国际会议（ICCV），IEEE（2017）5727-5736

43. Ghanem, B., Niebles, J.C., Snoek, C., Heilbron, F.C., Alwassel, H., Khrisna, R., Escorcia, V., Hata, K., Buch, S.: Activitynet challenge 2017 summary. arXiv preprint arXiv:1710.08011 (2017)
43.Ghanem，B.，Niebles，J.C.，Snoek，C.，Heilbron，F.C.，Alwassel，H.，Khrisna，R.，Escorcia，V.，Hata，K.，Buch，S。：Activitynet challenge 2017 summary。 arXiv preprint arXiv：1710.08011（2017）

44. Tran, D., Ray, J., Shou, Z., Chang, S.F., Paluri, M.: Convnet architecture search for spatiotemporal feature learning. arXiv preprint arXiv:1708.05038 (2017)
44. Tran，D.，Ray，J.，Shou，Z.，Chang，S.F.，Paluri，M。：Convnet architecture search for spatiotemporal feature learning。 arXiv preprint arXiv：1708.05038（2017）

45. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2014) 1725–1732
45. Karpathy，A.，Toderici，G.，Shetty，S.，Leung，T.，Sukthankar，R.，Fei-Fei，L。：具有卷积神经网络的大规模视频分类。在：IEEE计算机视觉和模式识别会议论文集。 （2014）1725-1732

46. Zhao, Y., Zhang, B., Wu, Z., Yang, S., Zhou, L., Yan, S., Wang, L., Xiong, Y., Lin, D., Qiao, Y., Tang, X.: Cuhk & ethz & siat submission to activitynet challenge 2017. arXiv preprint arXiv:1710.08011 (2017)
46. Zhao，Y.，Zhang，B.，Wu，Z.，Yang，S.，Zhou，L.，Yan，S.，Wang，L.，Xiong，Y.，Lin，D.，Qiao，Y。，Tang，X。：Cuhk＆ethz＆siat提交给2017年的活动网络挑战.arXiv preprint arXiv：1710.08011（2017）

47. Wang, R., Tao, D.: Uts at activitynet 2016. AcitivityNet Large Scale Activity Recognition Challenge 2016 (2016) 8
47. Wang，R.，Tao，D。：Uts at activitynet 2016.AcitivityNet大规模活动识别挑战2016（2016）8

48. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. arXiv preprint arXiv:1703.03329 (2017)
48. Wang，L.，Xiong，Y.，Lin，D.，Van Gool，L。：用于弱监督行动识别和检测的Untrimmednets。 arXiv preprint arXiv：1703.03329（2017）

49. Heilbron, F.C., Barrios, W., Escorcia, V., Ghanem, B.: Scc: Semantic context cascade for efficient action detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Volume 2. (2017)
49. Heilbron，F.C.，Barrios，W.，Escorcia，V.，Ghanem，B。：Scc：用于有效行动检测的语义上下文级联。在：IEEE计算机视觉和模式识别会议（CVPR）。第2卷（2017）

50. Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.F.: Cdc: Convolutional-deconvolutional networks for precise temporal action localization in untrimmed videos. arXiv preprint arXiv:1703.01515 (2017)
50. Shou，Z.，Chan，J.，Zareian，A.，Miyazawa，K.，Chang，S.F。：Cdc：Convolutional-deconvolutional networks，用于在未修剪的视频中进行精确的时间动作定位。 arXiv preprint arXiv：1703.01515（2017）

51. Xiong, Y., Zhao, Y., Wang, L., Lin, D., Tang, X.: A pursuit of temporal accuracy in general activity detection. arXiv preprint arXiv:1703.02716 (2017)
51. Xiong，Y.，Zhao，Y.，Wang，L.，Lin，D.，Tang，X。：追求一般活动检测的时间准确性。 arXiv preprint arXiv：1703.02716（2017）

52. Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. arXiv preprint arXiv:1703.07814 (2017)
52. Xu，H.，Das，A.，Saenko，K。：R-c3d：用于时间活动检测的区域卷积3d网络。 arXiv preprint arXiv：1703.07814（2017）